HEALS

Health and Environment-wide Associations based on Large population Surveys

FP7-ENV-2013- 603946

http://www.heals-eu.eu/

D5.3 Report on best practices for -omics analysis performed on human cohorts

WP5 Omics, epigenetics and confirmatory in vitro analyses

Version 1

Lead beneficiary: 13 – TNO

Date: 21/02/2018

Nature: R - Report

Dissemination level: PU - Public

TABLE OF CONTENTS

Document Information

Grant Agreement Number	ENV-603946	**Acronym**	HEALS
Full title	Health and Environment-wide Associations based on Large population Surveys		
Project URL	http://www.heals-eu.eu/		
EU Project Officer	Tuomo Karjalainen,- Tuomo.KARJALAINEN@ec.europa.eu		

Deliverable	**Number**	5.3	**Title**	Report on best practices for -omics analysis performed on human cohorts
Work Package	**Number**	5	**Title**	Omics, epigenetics and confirmatory in vitro analyses

Delivery date	**Contractual**	M42	**Actual**	21/02/2018
Status	Draft ☐		Final ☒	
Nature	Demonstrator ☐	Report X	Prototype ☐	Other ☐
Dissemination level	Confidential ☐	Public X		

Author (Partners)	Bernice Schaddelee-Scholten, Andrew Povey, Ettore Meccia, Martien Caspers, Elwin Verheij, Kirsten Hertoghs, Michael Dickinson, Nafsika Papaioannou, Dimosthenis Sarigiannis, Rob Stierum			
Responsible Author	Rob Stierum		**Email**	rob.stierum@tno.nl
	Partner	TNO	**Phone**	088 866 17 79

1 Introduction

This deliverable aims to provide guidance on best practices for multi-omics analysis performed on human cohorts for the purpose of exploiting the human exposome to improve the association between environmental exposures and health. In this context this report aims to look at a higher level of detail into the analysis of human samples using high data content molecular biology and analytical techniques and in (pre-) processing of raw -omics data. The scope here is to identify and describe the best practices in these sample- and data processing steps that can effectively support the environment-wide and health association studies following the connectivity across different biological scales of organization and the hybrid untargeted-targeted approach that is characteristic of the HEALS paradigm.

Deliverable 5.2 provided an overview of available omics technologies including standard operating procedures for selected technologies that fall within the spectrum of techniques used in HEALS. In addition, D5.2 served to further select appropriate specific omics methodologies and project partners that could assist in the analysis of biomaterials from existing cohorts such as Repro-PL and Phime, which were prioritized for further analysis during the Ljubljana Workshop held in May 2014 as well as preparing initial protocols for novel cohorts (EXHES).

Deliverable D5.3 follows up on this by looking into more detail in following the analysis steps both considering the wet lab experimentation and (pre-) processing of data in relation to cohort studies.

Even though not applied to the cohort studies, some additionally methodology and suggestions for best practices on adductomics and proteomics is described here.

The PHIME and REPRO PL cohort studies were selected as existing cohorts candidate for molecular epidemiological studies available within the HEALS project. These studies are used as pilot cases for the application of multi-omics analysis in exposome studies, upon which to base the derivation of best practices that would then be followed for the multi-omics analysis of EXHES samples. The PHIME sample set is derived from a cohort of 675 mother and child samples. The general aim of this large study was to assess the impact of low levels of methyl mercury exposure through fish consumption during pregnancy on the neurodevelopment of children at 18 months. The Repro PL sample set is derived from a cohort of 400 mother and child urine and cord blood plasma samples. The study set up of both studies is described in detail in Deliverable 5.2.

2 Analyses

Fera has undertaken metabolomics analysis of *plasma* samples from both Repro PL and PHIME cohorts, using both Liquid Chromatography – High Resolution Mass Spectrometry (LC-HRMS) and Nuclear Magnetic Resonance Spectroscopy (NMR). The PHIME cohort was comprised of 2 subsets – 165 Mothers samples and 135 Child samples. These separate cohorts within PHIME were extracted and analysed as 2 separate metabolomics experiments to allow separate downstream data analysis. All samples were provided in tubes containing heparin anticoagulant. The REPRO-PL cohort provided 149 cord blood samples, with associated subject metadata. The best practices inferred from the LC-HRMS and NMR analyses of Repro-PL samples is described below in greater detail.

Partner AUTH has analyzed maternal *urine* by both NMR and high or ultra-performance liquid chromatography coupled to mass spectrometry (UPLC-MS/MS) platforms for both Repro PL and PHIME cohorts.

In addition Partner TNO helped with metabolome data preprocessing.

Partner GenomeScan (ServiceXS) has performed SNP profiling on DNA from Repro-PL and Phime, whereas Partner TNO performed the initial bioinformatics for SNP (preprocessing and SNP calling)

Partner ISS is currently analyzing the DNA methylation patterns from Repro-PL.

Partner University of Manchester has provided best practices for DNA adductomics.

2.1 Plasma samples

2.1.1 Sample preparation and analysis plan

149 samples of frozen cord plasma were received into Fera in April 2016. Samples were assessed for analytical suitability from both cohort sets. Only samples with volumes >300 µl could be analyzed by NMR whereas LC-HRMS, as a more sensitive technique, enabled the analysis of volumes down to approximately 20µl. Samples were then randomly assigned (using www.random.org sequence generator) into 5 analytical "batches" for each cohort of approximately 30 samples each, with 2 batches prepared each day. Within each batch a pooled quality control sample (made up of equal aliquots of blood from all cohort samples) labelled "In House Reference (IHR)" was extracted alongside each sample. The purpose of this was to understand any variability that may occur due to batch to batch extraction processes. The total number of batches for each cohort set was 5. Alongside the IHR for each batch, a pooled Quality Control (QC) (again, made up of equal aliquots of blood from all cohort samples) was extracted in bulk at the beginning of the experiment for instrument QC purposes.

Sample preparation and instrumentation parameters are described in detail in the second deliverable report and are based on optimized procedures described in Bruce et al. (2008). In short, the plasma sample is submitted to a protein removal step before being extracted for metabolites in a non-targeted fashion using 1:1 methanol: water. After centrifugation, the supernatant is split for NMR and LC-HRMS analysis. The MS aliquot is further diluted and the NMR aliquot is evaporated and reconstituted with phosphate buffer / sodium azide containing the internal standard 1mM trimethylsilyl propanoic acid (TSP) before acquisition. No internal standards (ISs) were used in the LC-HRMS analysis. From past experiences within the metabolomics program within TNO, the addition of internal standards to samples can be considered.

2.1.2 LC-HRMS acquisition

The LC-HRMS analysis was split into two further experiments – positive and negative ion mode acquisition, this was to increase the coverage of potential measurable metabolites. The IHR was analysed at the beginning of every batch and pooled QC were analysed every 6 (Repro PL) or 8 (PHIME) samples across the whole experiment, as a tool to normalise the sample data (metabolite signals), if required, against any batch drift effects across the 5 batches. Furthermore, to increase signal stability the column was conditioned with at least 7 dummy samples (data is discarded) before the sample acquisition began. (See Figure 1 for an example partial sequence). The total approximate acquisition time for each cohort set was 140 hours per ionisation mode. Within this 140 hours the MS is stopped after every batch (approximately every 30 samples) to evaluate the calibration (and re-calibrate if necessary) to maintain 5ppm mass accuracy of the detector.

	File Name	Sample ID	Position
1	001	Water	A:1
2	002	Column conditioner	A:2
3	003	Column conditioner	A:2
4	004	Column conditioner	A:2
5	005	Column conditioner	A:2
6	006	Column conditioner	A:3
7	007	Column conditioner	A:3
8	008	Column conditioner	A:4
9	009	IHR_B1	A:4
10	010	QC 1	A:3
11	011	R14005	A:5
12	012	N21050	A:6
13	013	R12030	A:7
14	014	R12106	A:8
15	015	R12023	A:9
16	016	R12105	A:10
17	017	QC 2	A:3
18	018	R12027	A:11
19	019	R11164	A:12
20	020	R12109	A:13
21	021	R12131	A:14
22	022	N21175	A:15
23	023	R11128	A:16
24	024	QC 3	A:3
25	025	R11097	A:17
26	026	N21154	A:18
27	027	N21107	A:19
28	028	N21116	A:20
29	029	R12156	A:21
30	030	N21217	A:22
31	031	QC 4	A:2
32	032	R12039	A:23
33	033	N21119	A:24
34	034	N21125	A:25
35	035	N21169	A:26
36	036	N21157	A:27
37	037	N21213	A:28
38	038	QC 5	A:2

Figure 1. The start of the sequence for the Repro PL cohort LC-HRMS experiment in positive ion mode. Figure shows column conditioners, the inclusion of an extraction variability check (IHR) and a pooled QC analysed every 6 samples.

2.1.3 LC-HRMS data pre-processing

Methodology of the data processing steps are described in detail in the second periodic report. Data from both cohorts was analysed using the following workflow:

i) **Raw data evaluation***:* Each Total Ion Chromatogram (TIC) and overall file size was scrutinised for obvious outliers. These are potential data files where there has been an instrument failure (e.g. mis-injection) in acquisition or serious error in the sample preparation. For the Repro PL cohort set, in the negative ionisation mode experiment one of the data files for a sample had to be discarded due to an extremely rare occurrence where the MS instrument locked the acquired data file, rendering the file useless after acquisition.

ii) **Pooled QC and IHR evaluation:** To get a first evaluation of instrument signal drift and / or extraction variability across the 5 batches, each IHR and pooled QC analysed was assessed by taking the peak area, retention time and mass accuracy of a number of example metabolites across the experiment. These include metabolites with varying masses, chemistries and therefore retention times (RT) in order to get a representative understanding of how the system is performing across the analytical run. Relative Standard Deviations (RSD) in response ideally should be < 25% for peak areas, < 0.5% for retention time and mass accuracy should be within +- 5ppm of the theoretical accurate mass at all times. Table 1 below provides a summary of a set of example metabolite signals (peak areas) for all IHR and pooled QC's analysed across the Repro PL positive ion mode experiment.

Table 1. % RSD peak areas for example metabolites across pooled QC and IHR data files for Repro PL positive ionisation experiment.

Metabolite	*m/z*, M+H	Retention Time	% RSD peak area Pooled QC (n=30)	% RSD peak area IHR (n=5)
Threonine	120.06551	1.9	14	15
Tryptophan	205.09714	10.4	16	23
Progesterone	315.23184	18.6	14	15
alpha - Linolenic acid	279.23184	18.5	19	24
Indole	118.06512	10.4	17	25
Caffeine	195.08764	11.1	15	20

iii) **Data alignment:** As described in Goodacre et al. 2007, large LC-MS metabolomics datasets such as the set acquired for the cohort samples should align all retention time / masses combinations (potential metabolite signals) across all the samples acquired in each experiment. This is to correct for any potential RT drift during the long analytical run. The alignment ensures any further data processing and comparison across samples is taking the relevant mass/RT responses for each sample, therefore maintaining accuracy of any further data analysis.

The cohort data sets were aligned using the commercial metabolomics software Progenesis QI from Waters Corporation. The software automatically assigned which data file was the most relevant to

align all others against and provided a "score" out of 100 for alignment of each of the data files, which was manually checked.

iv) **Peak picking:** Effective peak picking with reliable algorithms is crucial for accuracy of the number of signals (potential metabolites) taken forward for further analysis, with as few false negatives and false positives as possible. For the cohort data, metabolite signals were "picked" using the algorithms of Progenesis QI. These were masses (*m/z's*) at given retention times, resulting in an overall table of masses, RT and peak area / height (abundance) for every signal. A signal abundance threshold cut off was not used in this process but a low mass threshold was applied of < 80 *m/z*.

v) **Deconvolution:** After peak picking, certain metabolites had more than one mass associated with their signal. This is where different ionisation "adducts" such as +Na, +NH4, +K, etc. all provide different signals (masses) for the same metabolite. There can also be multiple signals detected for the isotopic peak of certain compounds. In the cohort data, Progenesis QI was used to look for groupings of signals to deconvolute a number of masses into one, thus reducing the volume of the data by representing each metabolite by just the one signal. The deconvolution of masses was then visually scrutinised to check accuracy of the approach. Figure 2 below summaries the deconvolution for an example metabolite.

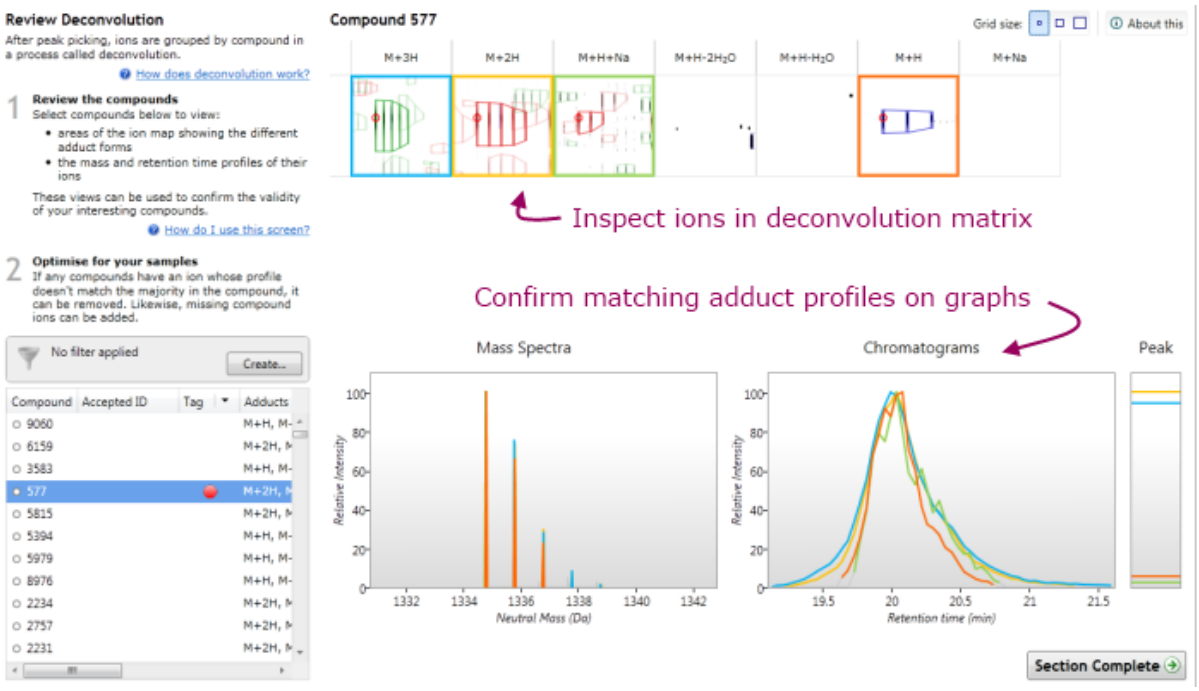


Figure 2. Example of deconvolution by Progenesis QI of a set of masses for a single metabolite. Example taken from http://www.nonlinear.com/progenesis/qi/v2.3/faq/reviewing-deconvolution.aspx

vi) **Batch correction:** As each cohort experiment was undertaken over 5 days using 5 extraction batches it was found that there was noticeable batch bias over the course of each analytical run. An example of this is shown for the Repro PL negative ion mode data set in Figure 3 below. This can be corrected using the pooled QC data or using a total signal approach for all samples as described in (Rusilowicz et al. 2016). Figure 4 shows a PCA of the same Repro PL negative ion data set after correction using the pooled QC. Batch correction should be undertaken before any data normalisation.

vii) **Data normalisation:** As described in van den Berg (2006), metabolomics data should be normalised or "cleaned" into a different scale in order to reduce the influence of measurement noise. Data acquired by LC-HRMS for both cohorts was normalised using Progenesis QI's scalar factor. The

software uses ratiometric data along with the mean and median raw data deviation to determine the scalar factor to be used for the normalisation. This approached is described in more detail in [1].

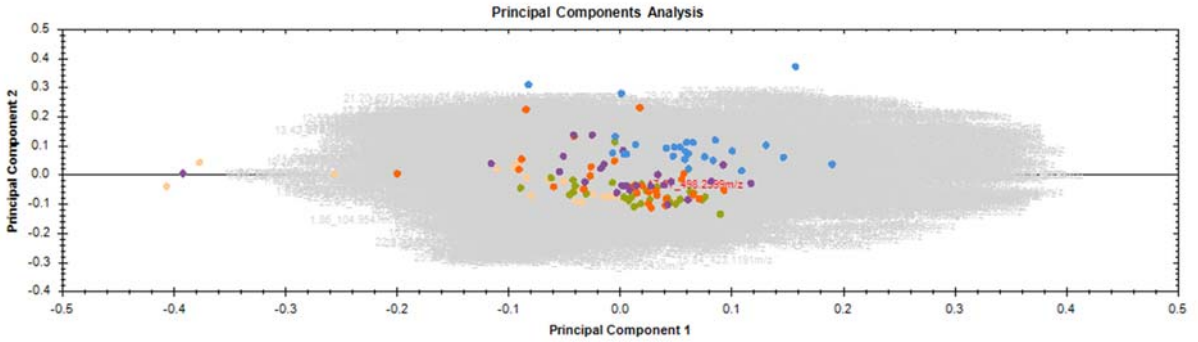


Figure 3. Principal Components Analysis (PCA) of all sample data from the Repro PL negative ion LC-HRMS data set. Colours indicate 5 sample batches , samples can be seen to partially group according to batch indicating a degree of batch to batch bias within the data set.

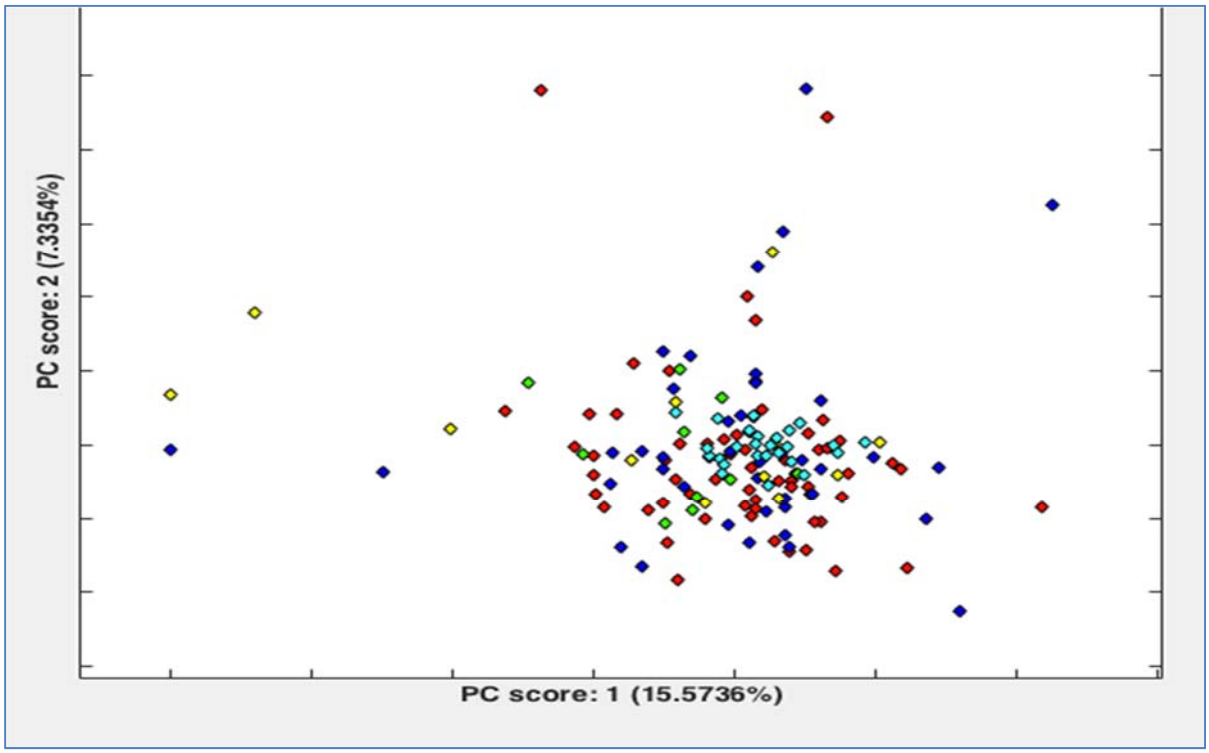


Figure 4. Principal Components Analysis (PCA) of all sample data from the Repro PL negative ion LC-HRMS data set after batch correction using pooled QC. Colours indicate 5 sample batches, samples do not group according to batch.

[1] NonlinearDynamics, in http://www.nonlinear.com/progenesis/qi/v2.3/faq/how-normalisation-works.aspx. (2017).

vi) **Metabolite annotation:** All signal data from all cohort analyses was assigned a potential metabolite annotation using both the freely available Human Metabolite Database (HMDB, www.hmdb.ca) and the Metlin Database (metlin.scripps.edu). This annotation included the common name, empirical formulae and unique CAS (Chemical Abstracts Service) registry number. It is important to note that annotations are purely derived from the accurate mass only and not any other analytical data, so they do not constitute or represent a full identification, and therefore metabolites should still be considered as unknowns. This means the annotations are tentative or a "level 4", as described in Salek et al. (2013).

2.1.4 NMR acquisition

Instrumentation parameters for the cohort experiments are described in detail in the second deliverable report (D5.2). Spectra were acquired at a central frequency of 500.1323505 MHz, using on-resonance pre-saturation to suppress the intensity of the water signal, followed by a 1D NOESY pulse sequence with irradiation of the residual water signal during the mixing time (200 ms). An observation pulse length of 10.0 µs and a delay between transients of 3 s were used. 65536 complex data points were acquired with a spectral width of 10400 Hz (corresponding to a chemical shift range of 14.0019 parts per million, ppm), giving an acquisition time of 4.679 s. Eight unrecorded (dummy) transients and 512 acquisition transients (scans) were used, giving a total experiment time of approximately 67 minutes per sample.

The IHR for each batch plus a single pooled QC was analysed with each batch of samples to monitor variation as a result of the sample preparation procedure.

2.1.5 NMR data pre-processing

Methodology of the data processing steps are described in detail in the second periodic report. Data from both cohorts was analysed using the following workflow:

i) **Free Induction Decay:** All cohort raw data were pre-processed using FELIX software from Accelrys. As described in McKenzie et al. (2017) - Various mathematical (window) functions can be applied to the free induction decay (FID), before it undergoes Fourier transformation (FT) to yield the spectrum. The application of certain window functions can dramatically increase the quality of the spectrum, yet generally there is a trade-off between the resolution between peaks and the signal-to-noise ratio (SNR).

Exponential (line broadening) functions were applied to the FID, which effectively weight the decay; more emphasis is placed where the time-domain SNR is greatest and less is placed at the tail end of the decay where the SNR is smaller. The function also forces the FID to decay to zero, which avoids the introduction of FT-induced artefacts. This function serves to increase the SNR of the spectrum.

ii) **Baseline Correction:** As described in McKenzie et al. (2017) - Definition of the spectral baseline is of paramount importance in the analysis of complex mixtures. Low concentration metabolites in NMR spectra can produce peaks of similar magnitude to noise, and differentiating these from the baseline is crucial, especially when the baseline is not optimally flat. Cohort data was baseline corrected using polynomial baseline correction applied in the Felix NMR (Accelrys Inc.) software.

The benefit of the method is that is does not require the identification of data points as either noise or signal, and is thus beneficial for application to spectra of complex mixtures, where a high density of signals would otherwise preclude the easy identification of noise. The method also imparts no rigidity by assuming the baseline to be of polynomial form.

iii) **Alignment:** As described in McKenzie et al. (2017) - Most spectral resonances do remain constant in spite of minor inhomogeneities that induce movement of other spectral peaks, yet such movements are generally uncorrelated. This non-correlation may result in spectra with some peaks shifting left, some shifting right, and some remaining intransigent. Thus, as there is no correlation, alignment is more than simply offsetting whole spectra; instead, only individual sections need to be manipulated, which are generally only a small proportion of the spectral width.

Cohort data was aligned using the chemical shift of the internal standard TSP, which has a resonance at 0 ppm.

iv) **Signal (feature) extraction:** NMR spectra can consist of tens of thousands of data points, yet performing multivariate analysis with so many signals is inefficient, as so many of the variables are correlated as spectral peaks. Introducing methods to retain all of the information whilst at the same time reducing the number of signals is a key step before further chemoinformatic analysis.

^{1}H NMR spectroscopic cohort data were "binned" using Metabolab, an in-house written GUI for the statistical software package MATLAB (Mathworks), using an adaptive binning algorithm based on the undecimated wavelet transform (Davis et al. 2007). This algorithm corrected minor variations in the chemical shift, performed feature selection, removed regions of the NMR spectra that only contain noise and performed data reduction.

v) **Pooled QC and IHR evaluation:** From the cohort data, to get an initial evaluation of instrument signal drift and / or extraction variability across the 5 batches, plus to check for obvious outliers, all binned data were scrutinised on a PCA. This is shown in Figure 5 for the NMR Repro PL data set.

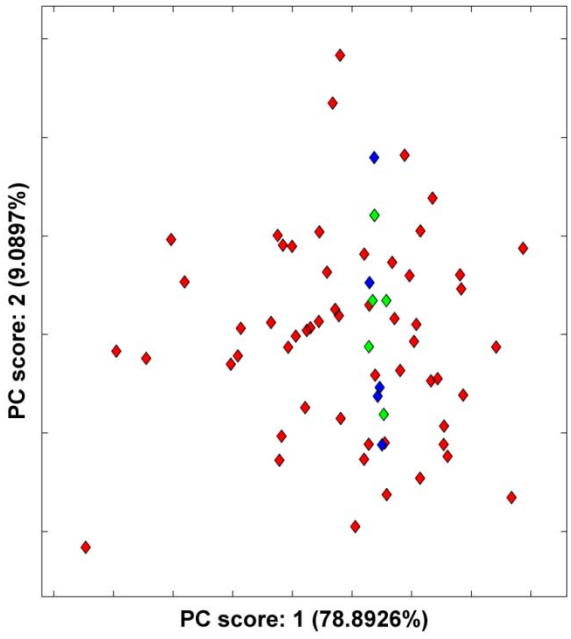


Figure 5. Principal Components Analysis (PCA) of all sample data derived from cord blood from the Repro PL NMR data set. Red = all samples, green = pooled extract QC sample, blue = IHR sample. The data does not show obvious outliers and all QC are satisfactorily clustered in the centre of the PCA.

Similar analyses were performed for Phime.

vi) **Metabolite annotation:**

On ^{1}H NMR data, such as acquired for the cohort analyses, spectra are heavily crowded which makes metabolite annotation challenging. As described in McKenzie et al. (2017), software packages and routines are capable of analysing and assigning one-dimensional spectra (e.g. MetaboMiner from Xia et al. (2008) , however, constituents of crowded spectra may not be solely differentiable based on peak shifts, and may require peak multiplicities to sure up any analysis.

Cohort data was subject to putative metabolite assignment using a combination of in-house databases, the MMCD (Madison Metabolomics Consortium Database) and the HMDB (Human Metabolome Database).

2.1.6 Additional recommendations on collecting blood plasma samples

Although proteomics analyses were not performed on the human samples, the next paragraphs describe the recommended protocol to collect blood plasma samples and explain the reasons to choose plasma instead of serum samples. Within HEALS it was not feasible to perform large scale proteome analysis on existing cohorts Repro-PL and Phime, as it was anticipated, with potential prolonged storage at room temperature at time of sample collection, samples were incompatible for proteome analysis.

Blood is one the preferred diagnostic materials in humans due to its accessibility and the vast medical laboratory infrastructure already in place for its analysis; and therefore, a logical biofluid to use for biomarker discovery.

Serum is very similar to plasma, expect for the absence of coagulated blood proteins. While one of the most frequently analyzed specimens, the generation of serum is time consuming and associated with proteolysis, coagulation cascade, and complement system. Thus, it can be expected that the reproducibility of serum samples is comparatively low. On the contrary, the sampling of plasma is less time consuming and more reproducible from a biochemical standpoint. Moreover, the Human Proteome Organization (HUPO) launched in 2002 the HUPO Plasma Proteome Project (HPPP), aiming to mine the serum/plasma from a proteomics point of view (Omenn et al. 2005). Using a well-scrutinized multicenter analyses of both serum and plasma, the HPPP concluded that plasma may be the preferred blood specimen for proteomics-based analysis providing better reliability and reproducibility (Rai et al. 2005)

Sample collection: Human blood plasma can be obtained by venipuncture (in healthy donors) or using peripheral access (in patients). Samples should be collected in K_2EDTA treated tubes (BD Biosciences). Blood samples have to be centrifuged twice at 1,200 × g for 10 min to pellet cellular elements, within 10 to 20 min of sample collection. Then, supernatant plasma has to be stored at -80°C until use.

Sample preparation: In the human plasma, the 22 most abundant proteins are responsible for ~ 99% of the bulk mass of the total proteins. So thousands of other proteins contribute for only ~ 1% of the protein mass(Anderson and Anderson, 2002). Hence, removal of highly abundant proteins during the plasma sample preparation process is normally performed in order to provide higher sensitivity for achieving broader proteome coverage. Immunoaffinity depletion of highly abundant proteins can be performed using pre-packed liquid chromatography columns (IgY14 Seppro®,Sigma; or similar), designed to remove Albumin, IgG, IgM, IgA, Transferrin, Haptoglobulin, alpha1-antitrypsin, alpha-1-acid glycoprotein, Fibrinogen isoforms, alpha-2 macroglobulin, Complement C3, and Apolipoproteins AI, AII and B.

After depletion, lower abundant proteins are denatured, reduced, alkylated and trypsin digested. The tryptic digestion are then analyzed by high accurate and sensitive mass spectrometry coupled to liquid

chromatography systems (LC-MS/MS) in order to identify proteins and quantify their level of abundance in the different samples analyzed.

Proteomic studies aim to identify proteins that are consistently changing abundance under different conditions. Therefore data analyzes is performed to estimate if there are significant differences in protein expression levels between healthy and pathological or exposure conditions, which would indicate a protein candidate for biomarker.

2.2 Urine samples

Urine from pregnant mothers in their third trimester was analysed by AUTH by both UPLC-MS and NMR analysis. As described in detail below, the final operating protocols followed were different from those described in D5.2. This need was something that emerged from the use of different instrumentation and the optimization of the operating conditions based on the instrument that was ultimately used (Annex 3, Table 1). Similar workflows were followed as described above for the blood analysis however software used was different between FERA and AUTH.

Note that in the case of the PHIME sample analysis it was decided to split the experiment in two batches. One batch for maternal urine and one batch for child's urine. All the protocols for PHIME urine samples were the same as in the case of REPRO PL project as the same instrumentation has been used for the analysis, and all conditions have been optimized in the analysis of the REPRO_PL samples.

2.2.1 LC-MS

2.2.1.1 Data acquisition

For LC-MS analysis, all urine samples were thawed in controlled conditions according to Nature protocols (Theodoridis, et al., 2012; Want, et al., 2010). A quantity of 600 µl urine samples was centrifuged (10000 rpm for 10 min) to remove all precipitates. After that, 500 µl of each clear urine sample was placed on autosampler vials and diluted with 1000 µl of LC-MS water, ready for analysis. Furthermore, a quantity of 50 µl was used for the preparation of the pooled QC sample following the same procedure as the samples. ThermoFisher Scientific (Bremen, Germany) model LTQ Orbitrap Discovery MS with a resolution 30,000 system was used. Sample analysis for urine samples was carried out in positive and negative ion modes. The mass scanning range was set between 50–1000 m/z and the capillary temperature at 320°C. The flow rate for nitrogen sheath gas and auxiliary gas was 40 L/min and 8 L/min respectively. Spray voltage was 4.5 kV, with the LC–MS run in a gradient mode with two solvents. Mobile Phase A: 100% high-grade water with 0.1% formic acid and mobile phase B: 100% acetonitrile with 0.1% formic acid or 100% methanol with 0.1% formic acid: the flow rate was 500 µl/min for the urine samples analysis in both positive and negative mode. Chromatographic separations were achieved using an Acquality UPLC HSS T3 column (100 x 2.1 mm, 1.8 µm, Waters, Milford, MA, USA) maintained at a constant temperature of 40 °C. The following gradient was used for both positive and negative mode: 1% B at 0 min, 1% B at 1 min, 15% B at 3 min, 50% B at 6 min, 95% B at 9 min, 95% B at 10 min, 1% B at 10.1 min and 1% B at 14 min. Two blank samples were injected before the starting QC of each run for checking the column, and ten QC pooled samples were injected for the baseline stabilization of the LC–MS system. All samples in the positive mode were run as one batch. Also in negative mode the samples were run as one batch. In the urine sequence, the pooled QC sample repeated every ten samples, and every twenty samples a solvent blank sample was repeated. It is highly recommended that every 100 runs the probe and the capillary must be handled and cleaned properly (Annex 3, table 2).

2.2.1.2 Data pre-processing

The main goal in spectral data processing is to correctly arrange the huge amount of raw data generated by the MS files into a 2D matrix in which each column represents one sample (object) and each row one ion detected by the MS (variable). Variables are expressed by the combination of three characteristics, two that define the identity of the ion, that is m/z and RT, and the ion intensity measured for each sample.

Data was acquired using the ThermoFisher Scientific model LTQ Orbitrap Discovery MS, and for data generated by Non-Agilent Chromatography system (non-Agilent data), spectral data processing was performed using the MZMine v.2.21 open-software (Katajamaa, et al., 2006). Raw data generated from negative and positive ionization are treated as two different experiments. Spectral data processing included the following steps:

2.2.1.1.1 Baseline Correction

Baseline correction was used to remove low-frequency artifacts and differences between samples that are generated by experimental and instrumental variation. After this, the application of high-frequency filters may be necessary to remove the electronic noise present in the data that is generated by the measurement equipment (Alonso, et al., 2015). The use of smoothing, for which the main purpose is to remove high-frequency noise from the chromatograms prior to deconvolution, is highly recommended. The decision of how many mean values will be taken into consideration is crucial and requires testing.

2.2.1.1.2 Peak (or Mass) detection

Peak detection algorithms analyze each sample spectrum independently. The different metabolites are identified using one or multiple detection thresholds. These thresholds are applied to different parameters such as the signal-to-noise ratio, the intensity or the area of each peak (Alonso, et al., 2015).

The Mass detection module generates a list of masses (ions) for each scan in the raw data file. Several algorithms are provided for this step. The choice of the optimal algorithm depends on the raw data characteristics (mass resolution, mass precision, peak shape, noise). For example, if the raw data is already centroid[2], only Centroid mass detector algorithm can be used, which simply assumes that each signal above given noise level is a detected ion.

Moreover, the choice of noise level value depends from the detected peaks that do not represent metabolites and must be excluded from the further analysis. So, it depends from the detected peak of a metabolite with the minimum height.

2.2.1.1.3 Chromatogram Builder

The chromatogram builder takes the mass lists generated by the mass detection step and builds a chromatogram for each mass that can be detected continuously by the scans. At the end, the

[2] Depending on the instrument, MS data can be recorder in two different modes. The generated files from a Fullscan in orbitrap, for example, are in mzXML, and these files can be acquired using profile or centroid mode. In centroid mode, each ion is represented as a discrete m/z, intensity pair, while in profile mode the ions are represented by peaks each containing a collection of points. We acquire directly in centroid mode, which give a smaller file and requires less processing time when you use MZmine software.

chromatogram may have more than one peak and that is why the chromatogram deconvolution is a necessary step.

In this step, m/z tolerance and ppm must be defined. m/z tolerance refers to the absolute difference given normally in Da, in amu or u (unified atomic mass unit). The ppm is the relative tolerance. MZmine calculates the range of tolerance with the maximum of the absolute and relative tolerances.

$$ppm = \frac{observed\ mass - calculated\ mass}{calculated\ mass} \cdot 10^6$$

For example, when caffeine is used in a typical experiment the following applies:

$$ppm = \frac{195.0866 - 195.0877}{195.0866} \cdot 10^6$$

$$tolerance \sim 5ppm$$

Usually chosen values for TOF are 10 to 15 ppm from 0.003 to 0.004 m/z and for Orbitrap 5 ppm and m/z below 0.0015. In order to define its value, caffeine or reserpine spectrum from the analysis must be used as a reference. This is the reason why reference samples in the beginning of each experiment are used and sometimes runs are repeated during the sequence in non-specific time. Then the observed mass of the reference sample is related to the calculated one in order to calculate the relative tolerance needed for MZmine as shown above.

2.2.1.1.4 Chromatogram Deconvolution

Peak overlap is a common problem in MS-based studies. Overlapping peaks are treated as one. To attempt to solve this problem, chromatogram deconvolution methods have been developed. (Alonso, et al., 2015).The most common choice with the AUTH lab on urinary samples to perform deconvolution is an algorithm called Wavelets (XCMS) because it is faster and more suitable for data sets with little noise. This method uses wavelets to detect peaks within a chromatogram. A series of wavelets of different scales is convolved with the chromatogram. Local maxima in the convolution results determine the locations of possible peaks. When these candidate peak locations co-occur at multiple scales than the scale with the strongest response indicates peak width. Given the candidate peak locations and scales, peaks can then be reconstructed from the original chromatogram. Local maxima should be used instead of Wavelets (XCMS) only in the cases where a limit number of metabolomics are being identified through the XCMS deconvolution.

2.2.1.1.5 Isotopes

Isotope grouper module attempts to find those peaks in a peak list which form an isotope pattern. The difference between neighboring isotopes is a single neutron or 1.008665 Da, but part of this mass is consumed as a binding energy to other nucleons. This small difference may become significant with high-resolution MS data. The actual mass difference between isotopes depends on the chemical formula of the molecule. Since MZmine does not know the formula at the time of deisotoping, it assumes the default distance of ~ 1.003 Da, with user-defined tolerance the m/z tolerance parameter.

For small molecular weight compounds with monotonically decreasing isotope pattern, the most intense isotope should be representative. For high molecular weight peptides, the lowest m/z peptides, the lowest m/z isotope may be representative (Cañaveras, 2015).

2.2.1.1.6 Peak alignment

Peak alignment is one of the main processing steps in metabolomics studies involving multiple samples. The position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts that are usually introduced by differences in the chemical environment of the samples, like ionic strength, pH, or protein content. In MS-studies peak shifts are observed across the retention time axis (Alonso, et al., 2015).

The recommended algorithm for peak alignment is the RANSAC, which is a nondeterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed because can estimate parameters of a mathematical model from a set of observed data that contains outliers.

2.2.1.1.7 Gap Filling

Following alignment, the resulting peak list may contain missing peaks as a product of a deficient peak detection or a mistake in the alignment of different peak lists. The fact that one peak is missing after the alignment does not imply that the peak does not exist. In most cases, it is present but was undetected by the previous algorithms.

After gap filling, the user can export m/z values, retention times, and peak area for each identified peak in CSV format.

2.2.1.3 Metabolites Annotation

The most crucial step for downstream bioinformatics analysis and one of the biggest challenge is the annotation of metabolites. The mass-to-charge ratio (m/z) value of a molecular ion of interest is searched against metabolite database(s). The metabolites having molecular weights within a specified tolerance to the query m/z value are retrieved from the databases as putative identifications. These putative identifications serve as a foundation for further metabolite verification. It is important to use multiple sources, in order to induce the possibility of missing information. The use of both METLIN and HMDB databases for metabolite identification is highly recommended.

The Human Metabolome Database (HMDB) is a freely available electronic database containing 74,507 metabolite entries including both water-soluble and lipid soluble metabolites. METLIN database, on the other hand, includes masses, chemical formulas and structural detail for over 15,000 endogenous and exogenous metabolites and di- and tri-peptides. The database also facilitates high-volume research with automatic searches using mass lists.

In addition to searching for m/z values only, the ion annotation information can be used to aid the mass-based search. Ion annotation groups the ions originating from the same metabolite together and annotates them as adducts/isotopes/in-source fragments. R package CAMERA (Collection of Algorithms for metabolite profile Annotation) was previously developed for ion annotation by Kuhl etc. (Carsten Kuhl etc. CAMERA: Collection of annotation related methods for mass spectrometry data. R package version 1.10.0.). Using the ion annotation information, the appropriate mass values of ions can be calculated. Then the calculated mass values are searched against databases. This approach is expected to improve the accuracy of metabolite identification.

Metabolites annotation was performed using mass-based metabolite search simultaneously against the two major metabolite databases: Human Metabolome DataBase (HMDB), and Metlin. The search results from these databases are integrated into a uniformly and non-redundant format based on IUPAC International Chemical Identifier (InChI) key. Chemical identifier information is provided by the software for effective reference to metabolites. Cross-referencing across multiple databases is performed when a particular identifier type is missing from a database. The comprehensive list of chemical identifiers includes PubChem Compound ID (CID), PubChem Substance ID (SID), HMDB ID,

KEGG ID, InChI string, and InChI key. MetaboSearch performs a mass-based search using a given list of m/z values. In addition, it can utilize ion annotation information for improved metabolite identification, as long as the ion annotation information is provided according to the CAMERA output format. The query of the exact mass of the detected features against on line databases within a certain mass range (±10 ppm) (Zhou, et al., 2012).

Another recommendation regarding metabolites, in the case that metabolic pathway analysis is the next step, is to limit the query to endogenous metabolites while performing compound identification against Metlin and/or HMDB database. This would increase the number pathways matched for LCMS data.

Also, sometimes the compounds that have been annotated with the CAS ID annotation may be synthetic/drug based compounds. These compounds must be noted and were excluded from the final metabolite list if there is no evidence that the participants used to consume them.

The limitation of using LC-MS techniques for the untargeted metabolomic analysis in exposomic studies is the fact that above a specific number of samples the analysis cannot be performed in one batch. The tolerable number of samples that can be analyzed in one experiment depends on the instrument specifications. The problem with the separation of samples into smaller batches is that is almost impossible to perform the experiments under exact the same analytical conditions. The lack of repeatability makes the troubleshooting during bioinformatic analysis way more difficult than it has already been. If the experiment has to be performed in different batches, also the data should be analyzed (Mzmine, etc) in different batches using the specific conditions were used in each batch.

2.2.2 NMR

2.2.1.4 Data acquisition

In NMR analysis, urine samples were thawed and centrifuged (10000 rpm for 10 min) removing all the precipitates. Then, 500 μL of the supernatant was placed in clean Eppendorf mixed with 120 μL of buffer solution (Na_2HPO_4 0.2 M, NaH_2PO_4 0.3 M in 50% D_2O / 50% H_2O) containing 0.1% TSP-*d*4 (used as chemical shift reference (δH 0.00 ppm)). The samples were then vortexed for 1 min and placed in -4^oC for 7 min. Subsequently, the samples were centrifuged (10000 rpm for 10 min) and 550 μL of each was transferred to a 5 ml NMR tube. The final pH of the samples was 7.337.

Urine samples were acquired on a 600 MHz Varian spectrometer, using a spectrometer frequency of 599.938 MHz with an OneNMR Probe and a ProTune System (Agilent) using on-resonance pre-saturation to suppress the intensity of the water signal. Spectral size range covered from −2 to 10 ppm (spectral width 9615.4 Hz). Proton spectra were acquired with 128 scans with a relaxation delay of 2 s, acquisition time 4 s and pulse width 8.587 μs.

2.2.1.5 Spectral Data Processing and Metabolites Identification

In this study, spectral analysis proceeded using MestReNova (Mnova 11.0.3) (http://mestrelab.com/), while for the identification of metabolites it was deemed necessary to use in addition ChenoMx (http://www.chenomx.com /). Reason was also the direct connection to downstream usage in GeneSpring analysis software.

Spectrum analysis includes the following steps: After loading the spectra of all the samples, place one behind the other by using the command *superimposed*. Next step is to correct the position of the reference peak sample. In this study, the used reference was deuterium oxide (D_2O), due to the used

buffer. All reference peaks samples should be aligned one behind the other. The next step includes the correction signal intensity values (baseline correction). The most widespread algorithms for baseline correction is the Continuous Wavelet Derivative Transform (CWT) and the Smoother Whittaker (Carlos Cobas, et al., 2006; Jewison, et al., 2014). A measure of algorithm suitability is whether the algorithm creates a problem in the phase of the spectrum. There is the possibility that after baseline correction with the aid of an algorithm some peaks shift from the positive to the negative part of the axis of tension (y axis) or vice versa. This means that this specific algorithm is not suitable for the analysis. After that, spectrum phase must be checked and, if necessary, corrected. For phase correction, an automatic algorithm is preferable. A chromatogram resulting from an NMR untargeted metabolic analysis of a sample will contain more than 22000 variables. It is, therefore, necessary to reduce the volume of information to enable the investigator to end up to conclusions after further analysis. This reduction has taken place by grouping these variables (binning or bucketing). In more detail, the user split the x-axis into smaller regions, setting a value for the length of each region. It is common practice to choose length values less than 0.04 ppm. Also, all the peaks should belong to the same region. In each case, for the success of binning, the final image should be identical to that of the original spectrum (Smolinska, et al., 2012; Weljie, et al., 2006). Then the spectrum should be imported into the program ChenomX NMR Suite 8.2, where there will be the identification of the peaks. The identification of the peaks is based on compounds that exist in the library of the program and requires the definition of pH and concentration of TSP in the buffer that was added to the samples, since these parameters define the position of the peaks (Amiot, et al., 2015; Beaudry, et al., 2016). Finally, returning to the program MNOVA peak integration of TSP and metabolite, which has been identified previously, takes place and the peaks of metabolites are identified. In cases where a metabolite is characterized by multiple peaks and/or peaks in different areas of ppm, the area of all these peaks should be added to fill the corresponding cell on the sheet of import file to MPP.

2.2.1.6 Metabolomics data processing and data analysis results.

Metabolites identification revealed that the total number of putatively identified metabolites in urine samples analysis is quite bigger than the one that was revealed by the analysis of plasma samples. Also, in the case of plasma samples, the number of metabolites resulting from analysis in positive ionization is greater than the corresponding to negative, unlike urine samples. Possibly this might be due to the chemical composition of the above biological fluids, but also due to the fact that the urinary metabolites are the final products of metabolism and are characterized by higher sensitivity.

More specifically, 1715 out of 3239 potential metabolites that were identified from the analysis of urine samples in the negative ionization and 623 of the 1009 metabolites resulting from the analysis of urine samples in positive ionization, were totally characterized. Respectively, only 264 of the 267 potential metabolites resulting from the negative ionization of plasma samples and 380 of the 430 potential metabolites resulting from their positive ionization, were identified.

The fact that not all of the potential metabolites nor in the positive nor in negative ionization were annotated is probably due to the gaps of current databases of metabolites. Constantly adding information takes place and the bases are renewed.

Samples analysis using NMR spectroscopy, after data processing and metabolite annotation, resulted in the identification of 40 metabolites in urine samples and 49 in plasma samples.

Various sources were used to putative metabolite assignment: for LC-MS HMDB and Metlin, for NMR Chenomx

2.3 SNP analysis

The major goal of the SNP-analysis task is to select and apply methods for identification of interactions between Child-SNPs (of child cord blood) and parameters describing Child-Exposome (Environment/Mother/Father) and Child-End-points (eg. Health, Physiology, Cognition, Molecular markers). The SNP profiling pilot to REPRO-PL was not aimed at a complete GWAS (linking specific individual SNPs strongly to health endpoints) as the total number of subjects is too low, in particular for less frequent SNPs. Instead the focus was on possibilities of combining data from the metabolome and SNP analysis, to contribute to the understanding of the endotype, the internal mechanistic constitution, underlying the external phenotype. At the time of writing this report, the interpretation of the plasma metabolome data was still ongoing. Based upon the final pathway based outcome, a future interpretation of the SNP data in light of pathway perturbations is foreseen.

Concerning Repro-PL Cord blood was collected during child-delivery and named with the Mother/Child-pair identifier (MC-id), which was performed under the coordination of partner NOFER. Part of the blood was shipped to partner ISS for isolation of DNA and sent subsequently to the HEALS-partner performing the SNP-array analysis (GenomeScan, Leiden).After producing the SNP data TNO further analysed the dataset.

A similar logistical approach was followed for samples obtained from Phime.

For both Repro-PL and Phime, genomic DNA was isolated from buffy coats and subjected to SNP profiling. Attempts were also made by ISS to extract intact mRNA from the same samples, using a modification of the protocol by Hebels *et al.* , which consisted of mechanical separation using a saw, at low temperatures of the sample. However, this proved to be unsuccessful, as RNA was largely or fully degraded. Therefore the genetics analysis was further confined to SNP analysis only. Paragraph 2.3.4 further discusses a number of remarks on sample storage and planning for nucleic acids based omics.

Two platforms were initially considered for the implementation into HEALS exposome platform; eventually the Affymetrix Axiom® Biobank Genotyping Arrays and protocols was chosen: this is a whole genome SNP platform with GWAS markers selected from Axiom® Genomic Databases of SNPs and InDels.

2.3.1 SNP array analysis

M/C-id's of child cord-blood DNAs (received by GenomeScan from Study directors) were re-named with unique GenomeScan id's (GS-id), prepared and quality checked for SNP-array analysis. The M/C-id vs GS-id lists were shared with the REPRO PL and PHIME PIs and with TNO. The SNP-data and analysis report were shared on the HEALS/AUTH-server and delivered to HEALS partner TNO for further pre-processing and data-analysis with the other study data. Quality of 287 of 288 samples was classified OK (for 1 sample no SNP data obtained). One of the 287 samples showed 96.4% call-rate (just below 97% cut-off).

Remarks:

Pre-processing of SNP-analysis is a task involving specialized knowledge of the specific SNP-platform normally available at the SNP-analysis lab (GenomeScan). This task was transferred to TNO since GenomeScan no longer could support this service for the Affymetrix, Axiom_UKB_WCSG arrays (array processing was discontinued during the 1.5 year delay in sample receipt and is replaced by RNA sequencing). This resulted in an unexpectedly large effort for TNO to get familiar with SNP-pre-processing (see below). Additional expertise was provided by Affymetrix customer support to TNO.

From this a recommendation for best practices follows: ensure in projects concerned with long larger cohort studies continuity of both platform and downstream data processing capacity.

2.3.2 SNP-data pre-processing

Pre-processing of the SNP-array images (.cel files) was performed by TNO using the "AxiomAnalysisSuite"-Software (Affymetrix, freely available), -Database and -Default methodology-settings. Preprocessing resulted in successful genotyping results for 830115 SNPs and 287 samples (145 REPRO PL, 139 PHIME (3 are replicates of 1 sample "MH-K"), 3 Affy-controls).

2.3.3 Validation of gender

The main outcome of the SNP-analysis, child "SNP-gender", is the first SNP-analysis outcome that can be compared with the child "known-gender" as a check for the validity of the logistics and the analysis. The analysis of the HEALS REPRO-PL and PHIME samples resulted in some conflicts between SNP-derived and known child gender. Initially, upon the first analysis, 14 samples yielded known/SNP-gender conflicts (9 REPRO-PL, 5 PHIME samples) in the 3rd 96-well SNP-typing plate (and none in plates 1 and 2); see Annex 1. There were 5 female children with a male SNP-gender (4 Phime, 1 REPRO PL) and 9 male children with a female SNP-gender (8 REPRO-PL, 1 Phime). The samples with gender-conflicts had a high incidence of "manual samples transfer" into the SNP-typing plate 3 (REPRO PL: all 9, PHIME: 1 of 5).

In SNP-typing plate 1 and 2 there were 13 REPRO PL samples with "unknown" SNP-gender. Examination of these samples in a "chromXY-ratio-gender plot" (avgX vs avgY; Annex 2) in AxiomAnalysisSuite showed that most unknowns were plotted in the region of their expected gender but apparently did not reach the cutoff. Only 2 samples were plotted in an undecisive area. There were no unknowns with an obvious know/SNP-gender conflict.

The above mentioned 9 conflicting known/SNP-gender REPRO-PL samples are all plotted in very decisive areas of the below "chromXY-ratio-gender plot", indicating that the SNP-gender typing is ok, but more likely the identity of the samples is compromised.

Finally, we analysed whether the samples show signs of maternal contamination by analyzing the SNP calling data in the Axiom CNV (Copy Nr. Variant) analysis tool, yielding BAF-plots that show the genome-wide A/B/AB allele calling. Examination of the gender-conflict, unknown-gender and a set of correctly typed samples, all showed typical BAF-plot patterns for non-(maternal)-contaminated samples with a symmetric, 3 position patterns (for A/AB/B alleles; A or B at top or bottom; AB in the middle; Fig.6). No obvious bias for fuzzy versus clear BAF-plots was observed between the gender-conflict, unknown-gender and the correctly typed samples, indicating that maternal- or inter-sample contamination was not the cause of gender-conflicts and unknowns.

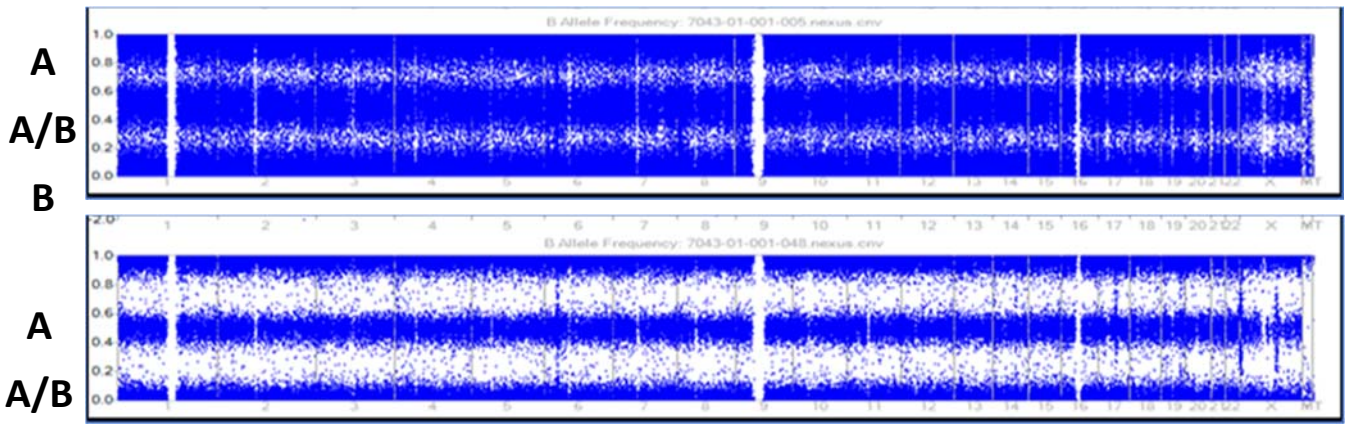


Figure 6: Two BAF plots showing the variation in quality of SNP-calling of 2 example samples. BAF plots axis: : Y = signal-ration for the 3 alleles (A, AB (middle), B); X = 800k SNPs of 23 chromosomes and the MT-DNA. Top and bottom plot show a fuzzy and a clear pattern, respectively.

From the above it was concluded that all gender conflict samples only occur in SNP-typing plate3 and that the conflicts seem to be related with sample-id confusion (logistics?) rather than with SNP-typing quality.

Therefore it was decided to use sample-cluster-analysis of all the SNP-typing data to solve the sample logistics of plate 3. Two samples were measured in each of the 3 plates as a quality control: Affy-Ctrl (a male Affymetrix DNA control sample) and MH-K (a female DNA sample from the Phime-study). As expected, in the SNP-cluster analysis 2 sub-clusters (each with 3 samples that cluster much closer together then all the other samples) were found indicating the 2x 3 replicates. One cluster contained the 2 Affy-Ctrl samples of plate 1 and 2 and another sample from plate 3, and the other cluster contained the 2 MH-K samples of plate 1 and 2 and another sample from plate 3. Finally, it became evident that an unintended conversion of a list of sample-plate-rows as a list of sample-plate-columns explained the unexpected position of the two control-replicates in plate3 (and of all other plate3 samples). After correction of the sample identities for plate 3 and a new SNP-analysis of all 3 plates (as one batch) correct SNP-based sample clustering of the 2 control triplicates and correct gender typing of all plate 3 samples was found.

From this, a recommendation towards best practices in multipartner exposome projects is to have access to central data infrastructure safeguarding sample tracking and annotation throughout to the level of assasy (in this case plate).

2.3.4 Remarks on RNA extraction

It was planned to extract RNA from the same samples for transcriptome analysis, but a number of problems were encountered (to be considered for future studies) that suggested to exclude RNA studies on the Repro_PL samples. They are discussed below:

1. Cord blood samples had been frozen as one single vial but DNA and RNA extraction procedures are not compatible. Frozen blood cells cannot be thawed for splitting the sample as the cell membranes will be disrupted, the sample will get too viscous to handle because of genomic DNA leakage and RNA will be degraded by free nucleases. To have good quality nucleic acids from blood samples cells, must be directly thawed, or ideally stored directly after collection, in the appropriate lysis buffer.

   We came to a mechanical procedure based on a rotary tool (Dremel) equipped with disposable cutting-off dishes for cutting aliquots from a unique criovial keeping it frozen. The procedure did work but was time and labour consuming and requested both equipment and manual skills unusual for a research lab. Directly storing the sample in several ready-to-use smaller aliquots seems to be a more appropriate solution. To this regard, it should be kept in mind that new omic approaches could be possible even if not planned at the time of sample collection.
2. Cut off frozen buffy coat aliquots should have been used for both genomic DNA and RNA extraction. A freshly frozen buffy coat from an adult donor was used as control sample. While DNA extraction was successful, neither standard procedures such as Qiagen and Norgen RNA extraction kits nor modified procedures based on direct lysis in 3.3 M Guanidine Thiocyanate/2-mercaptoethanol provided good or even sufficient quality RNA from the Repro_PL samples. As the fresh buffy coat did, as expected, this confirmed that there was a

problem with the Repro_PL samples and not with the extraction method. This was in accordance with our previous experience that blood samples for RNA extraction cannot be stored as such for too long. Tissues specimens like biopsies or cell pellets can be preserved and stored longer in RNA later or directly in RNA lysis buffer but this is not the case with blood samples. As for DNA, direct storage of ready-to-use sample aliquots added with lysis buffer as a preservative at the time of sample collection could probably have allowed good quality RNA extraction.

2.4 DNA methylation (epigenetics)

A growing number of studies is leading to a close association between exposure to environmental pollutants and changes in DNA methylation.

Endogenous factors like hormones, metabolites, cytokines and growth factors, as well as external factors like diet or drugs, are all able to dynamically modulate our unique genomic information. This will allow for cell and tissue type specificity in pluricellular organisms or for response to environmental factors by inducing epigenetic changes in our genome. A complex machinery of membrane and nuclear receptors, DNA and chromatin binding proteins acting as adaptors, readers and erasers, histone and DNA acetyl and methyl transferases together with specific transcription factors will translate the signals from external molecules, even at very low concentration, to the nucleus. The cooperative and self-reinforcing action of an increasing number of site specific chromatin modifications like histone acetylation, methylation, ubiquitination, sumoylation and DNA cytosine methylation will create an epigenetic code to tag and functionally define transcribed, non-transcribed, accessible and non-accessible regions of our genome. The resulting chromatin conformation can be reversible, irreversible, and heritable through the cell cycle or even transgenerational.

The same ability to sense and translate endogenous and external molecules and signals applies to many environmental contaminants able to interact with the epigenetic machinery as metals like cadmium or arsenic or hormone-like compounds. This interaction will likely disturb the proper cell regulation, interfering with cellular identity and functions, and will determine a broad spectrum of disfunctions at hormonal, metabolic, inflammatory level and last but not least participate to cancer predisposition or cancer development as a later event. Site specific, locus specific or global changes in histone post translational modifications and DNA methylation are possible exposure effects to be considered besides the more familiar genotoxic effects.

For what concerns population studies, DNA methylation is currently the only possible approach to associate exposure to the epigenome, as blood or blood derived buffy coat is typically collected during these studies, and good quality genomic DNA can be easily obtained from it. For this reason there is a growing interest in methylome analysis in population studies as the methylome can be considered both an effector and a biomarker of exposure.

At the time of D5.3 reporting, samples from the Repro_PL study were used to further explore the best practices for application of omics studies in human cohort studies. Cord blood samples for the Repro_PL cohort were collected and extracted as described in the SNP section. Aliquots of 500 ng genomic DNA have been used for methylome analysis in 128 samples.

The Illumina Infinium Methylation Epic Beadchip was chosen, as this array allows quantifying the levels of methyl-cytosine in more than 850.000 characterized CpG sites in the human genome.

CpG sites on the Epic array are distributed between CpG islands, CpG shores (sequences 2 kb upstream and downstream from CpG islands), and CpG shelves (sequences 2 kb upstream and downstream from shore regions) or are located outside these regions ("open sea").

According to a functional classification, they are located in proximal promoters (CpG sites located within 200 bp or 1500 bp upstream of transcription start sites, exon 1 and in 5´UTRs), 3´UTRs, and to gene bodies and intergenic regions. A detailed description of the array design can be found in Moran, Arribas and Esteller (2016).

Genomic DNAs from Repro_PL cord blood samples have been extracted, bsulphite converted with the EZ DNA Methylation kit from Zymo reserach and hybridyzed to the Infinium Epic methylation arrays.

The arrays have been scanned at the IIGM (Italian Institute for Genomic Medicine) Illumina platform in Turin (Italy) as previous projects in cooperation with the group where ongoing. At the time of this

reporting, the resulting idat output files have been shared on the HEALS/AUTH server for subsequent conversion to beta values, normalization and analysis.

2.5 Adductomics

DNA adduct analysis: In contrast to other –omic technologies, DNA adductomics is still in its infancy with optimisation still required for many of the basic methodological aspects (e.g. DNA extraction, DNA digestion) as it cannot be assumed that a procedure designed specifically for one type of adduct or adduct class is appropriate for another adduct type or class. Two different approaches being developed for both targeted and non-targeted approaches are described here. The first approach based upon Balbo et al, developed by UoM and FERA , is based upon the High Performance Liquid Chromatography (HPLC)–Tandem Quadrupole Mass Spectrometry (LC-MS/MS) and HPLC- High Resolution Mass Spectrometry (LC-HRMS) analysis of nucleoside digests of DNAs with the presence of adducts being confirmed by the neutral loss of the deoxyribose sugar moiety (-116 mass units). This approach potentially allows the targeted detection of a range of adducts through the use of adduct standards but also the non-targeted detection of other adducts (the standards of which are unavailable). The second approach being developed by the UoM uses a DNA repair protein, O^6-alkylguanine DNA-alkyltransferase to transfer alkyl groups from the O^6-position in guanine to the active site of protein which can then be digested with trypsin, enabling alkyl modified active site peptides to be detected by matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-ToF). This analysis is targeted specifically to one type of adduct known to be toxic, pro-mutagenic and carcinogenic.

Adductomics: In brief, DNA was extracted, digested to deoxyribonucleosides and analysed by both High Performance Liquid Chromatography – Tandem Quadrupole Mass Spectrometry (LC-MS/MS) and HPLC- High Resolution Mass Spectrometry (LC-HRMS) analytical platforms alongside a solvent blank, a deoxyribonucleoside standard mix (dA, dG, dC and dT) and O^6-methylguanine, N7-methylgunanine, 2-deoxyinosine and 2- deoxyuridine analytical standards.

DNA Digestion: DNA digestion was based upon the methodology of Badawi *et al*. Briefly, DNA was incubated with DNAse I, snake venom phosphodiesterase, and Alkaline phosphatase and 2'-deoxycoformycin for 16hr at 37°C in a Tris-HCL at pH7. Digested samples were then applied to a pre-conditioned C18 SPE cartridge before washing with 10% methanol and elution with 1ml 100% methanol. Eluents were evaporated to almost dryness in a centrifugal evaporator before reconstitution in 10% methanol in water.

Liquid Chromatography - High Resolution Mass Spectrometry (LC-HRMS) conditions: Liquid Chromatography was undertaken on an Accela LC system (Thermo Scientific). The analytical column used was a C18 ACE AQ with dimensions of 3 x 150 mm, 3 µm (Advanced Chromatography Technologies Ltd.). Mobile phase A (MPA) was ultra-pure water and mobile phase B (MPB) was methanol. A linear gradient elution was applied over 25 minutes from 90% MPA to 90% MPB. The gradient was then held for 10 minutes at 90% MPB before reequilibration with 90% MPA for a further 7 minutes. The LC flow rate was 0.4 mL min-1 and the column temperature was 25°C. Sample injection volume was 5 µL. High Resolution Mass Spectrometry was undertaken on an Exactive Orbitrap (Thermo Scientific). Analysis was in full scan single MS mode only, scanning between *m/z* 50 – 1000 at a resolution of 50,000 (at *m/z* 250) in positive ionisation mode. A heated MS ion electrospray source was used, set at 400◦C, source voltage at 3.5 Kv, sheath gas at 51 and auxiliary gas at 13 (both nitrogen, arbitrary units). Data was assessed using Xcalibur software (Thermo Scientific).

Liquid Chromatography – Tandem Quadrupole Mass Spectrometry (UPLC-MS/MS) conditions: Liquid chromatography conditions were identical to the conditions above for the LC-HRMS analysis, apart

from the LC was undertaken on an Acquity LC system from Waters Corporation. This system has a lower void volume to the Accela LC system resulting in differing (shorter) retention times by approximately 1 minute. Tandem Quadrupole Mass Spectrometry (MS/MS) was undertaken on a Xevo TQ-S (Waters Corporation). Analysis was by positive electrospray ionisation with multi reaction monitoring (MRM) of all analytes of interest. The ion source temperature was set to 150°C, with a capillary voltage of 1kV, cone voltage of 20V and desolvation temperature of 500°C. Desolvation gas was applied at 1000 L/hr, cone gas at 150 L/hour and nebuliser gas at 7 bar (all nitrogen). MRM transition conditions were optimised for the analytes and are described in Table 2. Data was assessed using Mass Lynx software (Waters Corporation).

Table 2. Optimised MS/MS multiple reaction monitoring conditions

Analyte (M+H, unless stated)	Transition	Cone Voltage (V)	Collision Energy (eV)
Deoxyguanosine (dG) (M+Na)	290>174	22	14
Deoxyadenosine (dA)	252>117	28	14
	252>136	28	14
Deoxycytidine (dC) (M+Na)	250>134	38	12
Deoxythymidine (dT) (M+Na)	265>139	30	20
	265>149	30	20
O^6-MeG / N^7-MeG	166>134	20	20
	166>149	20	18
2-deoxyuridine	251>139	32	10
	251>135	32	10
2-deoxyinosine	275>159	32	20
	275>117	32	10

Adduct detection: Using Xcalibur and Mass Lynx software, data from both methods (LC-MS/MS and LC-HRMS) were interrogated for the adducts detailed in Table 3. Peak areas or detected / non - detected were recorded as appropriate. Retention times detailed in this table are from the LC-HRMS set up and are marked as "unknown" if the adduct was not detected by either system. All MS/MS transitions were set up to monitor the neutral loss of the deoxyribose sugar moiety (-116 mass units) apart from O^6- MeG and N7-MeG which were derived and optimised from analytical standards.

Alkyl adductomics: In brief, following incubation with DNA or oligodeoxyribonucleotides containing specific O^6-alkylguanine adducts (ODNs), MGMT was digested with trypsin and the resulting tryptic peptides analysed by MALDI-Tof alongside chemically synthesised peptide standards (Senthong *et al* 2013).

MGMT treatment, recovery and digestion: MGMT was incubated with DNA or ODNs for up to 6hrs at 37°C in a Tris-EDTA-TCEP buffer. The DNA/protein solution was then added to a pre-washed Ni-coated

magnetic bead suspension and mixed at 4°C overnight. The beads were collected, washed and trypsin added and the beads incubated overnight at 37˚. Digestion was terminated by the addition of formic acid and the His-MGMT tryptic peptides were desalted and concentrated using Millipore® C18 Ziptips. Desalted tryptic peptides were eluted in 5 µL of 0.1% formic acid (FA) in 50% acetonitrile/water.

Table 3. List of adducts and their associated HRMS and MS/MS masses

Adduct	*m/z* M+H, HRMS	*m/z* M+H, MS/MS	Retention time (+- 0.3 minutes)
O^6-MeG	166.0723	166>149 , 166>134	6.5
N^7-MeG	166.0723	166>149 , 166>134	5
N^x-MedC	242.1135	242>126	4.5
N^x-ethyl-dC	256.1292	256>140	6.5
O^x-MedT	257.1132	257>141	6.5
N^x-MedT	257.1132	257>141	5
N^x-MedA	266.1248	266>150	8
N^x-ethyl-dT	271.1288	271>155	10.5
N^x-ethyl-dA	280.1404	280>164	5.5
O^6-MedG	282.1197	282>166	6.5
N^7-MedG	282.1197	282>166	5
1,N^6-etheno-dA	276.1091	276>160	6.5
8-oxo-dG	284.0989	284>168	4
N^2-ethyl-dG	296.1353	296>180	8
N^4-(4-OH-butyl-dC)	300.1554	300>184	unknown
O^4-(4-OH-butyl-T)	315.1551	315>199	11
α-OH-PdG	324.1302	324>208	10
S-γ-OH-Me-PdG	338.1459	338>222	5
N^2-(4-OH-butyl-dG)	340.1615	340>224	8
O^2-POB-dC	375.1663	375>259	7
O^2-POB-T	390.1660	390>274	10
O^2-POB-dG	415.1724	415>299	5
O^6-PHB-dG	417.1861	417>301	11
HNE-dG-1	424.2191	424>308	12
5-MeCDE-N^2-dA	544.2191	544>428	Unknown
5-MeCDE-N^2-dG	560.2140	560>444	Unknown
BPDE-N^2-dG	570.1983	570>454	Unknown

MALDI-ToF analysis and data acquisition: Each tryptic sample was spotted on a MALDI plate together with saturated α-cyano-4-hydroxycinnamic acid (Fluka, Buchs, Switzerland) matrix solution. The MALDI-ToF was calibrated by Peptide Calibration Standard (Bruker, Germany) or J67722 MALDI Certified Mass Spec Calibration Standard (Alfa Aesar, UK). Spectra were acquired over the m/z range 800–2300 using a Bruker (Germany) Ultraflex II operating at 30% laser intensity and 1000 laser shots per spectrum in Reflectron Positive Ion mode. Spectra were generated using FlexAnalysis (Bruker,

Germany) software and annotated using the MS-Digest tool of the protein prospector service (http://prospector.ucsf.edu: Figure 6). Data were analysed using Peptide Mass Fingerprint (PMF), a Mascot search tool on the Matrix Science website (http://www.matrixscience.com) with the SwissProt database at peptide mass tolerance of 0.5 Da with 2 allowed missed cleavages. A score greater than the Significance Level in Mascot and Expectation Values less than 0.05 were required. Mascot Expectation Values is the number of matches with equal or better scores that are expected to occur by chance alone. The lower the expectation value, the more significant is the score. S/N > 10 was required for identification of detected alkylated peptides ions. PAs of chosen peptides were measured by FlexAnalysis software (Bruker, Germany).

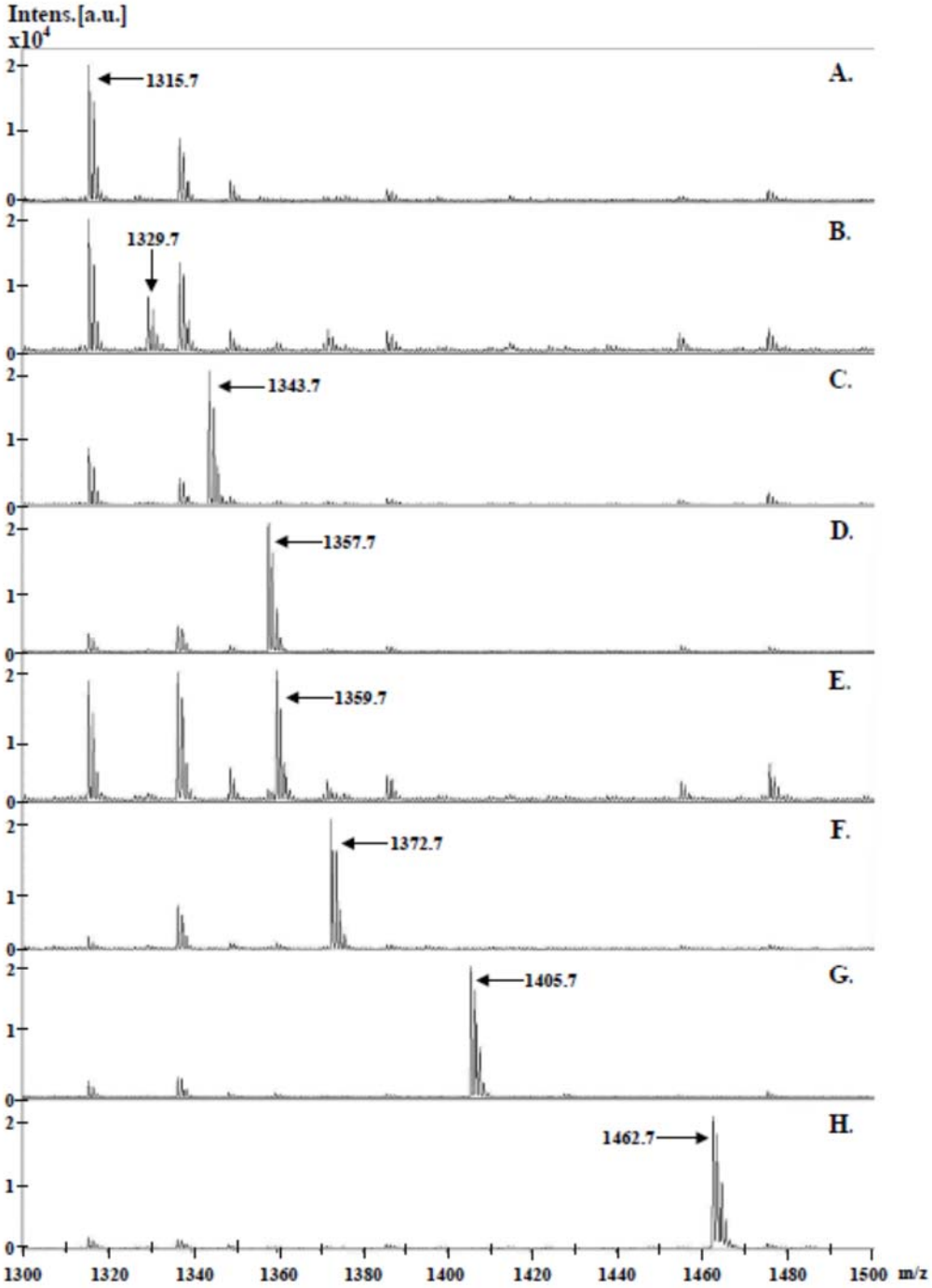


Figure 6. **MALD-ToF analysis of tryptic MGMT digests.** Locations and masses of the fragments expected are shown if alkyl group transfer occurred from ODNs containing A guanine; B O^6-methylguanine ; C O^6-ethylguanine; D O^6-propylguanine; E O^6-hydroxyethylguanine; F O^6-carboxymethylguanine; G O^6-benzylguanine and H O^6-pyridyl-oxo-butylguanine

3 Summary, recommendations and conclusions

Metabolomics

At the end of data processing, 510 out of 666 potential metabolites were annotated from the analysis of maternal urine samples in the negative ionization and 935 of the 514 metabolites were retrieved from the analysis of urine samples in the positive ionization mode. Respectively, 1263 of the 1754 potential metabolites resulting from the negative ionization of children's urine samples and 380 of the 430 potential metabolites resulting from their positive ionization, were identified.

Once acquired, all data was uploaded onto the AUTH FTP server for the bioinformatics team in HEALS to undertake subsequent statistical analysis.

Below, the main recommendations for best practices to apply different omics technologies in exposome studies are provided. This is based upon actual LCMS and NMR analysis and SNP analysis from urine and blood, obtained from the Repro-PL and PHIME cohorts.

Further, additional insightful considerations obtained from attempted transcriptome analysis on Repro-PL cord blood, as well as adductomics and proteomics under development and applied elsewhere within WP5 (in vitro) are provided as well.

LCMS

Data acquisition:

- One general recommendation referring to samples preparation for metabolomic analysis is to use for all samples only one brand of consumables. Note that the impurity of collection and storage tubes as well as filtration units should be tested prior to use.

Pre-processing data:

- Especially for large datasets it is recommended to look into batch correction (both between - and within batches). By using the pooled QC data or the total signal approach it is possible to correct for bias which could occur over the course of each analytical run.
-
- It is important to use multiple sources, for metabolites annotation in order to induce the possibility of missing information

NMR

Data acquisition:

- In order to maintain the repeatability during the experiment, it is necessary to add the same buffer to the samples, so buffer volume calculation before the preparation must be done very carefully, keeping in mind that some samples may be analyzed twice due to mistreatment.
- If a large number of samples is being studied (>10), spectral acquisition should be performed after sample randomization in order to avoid biasing results due to instrument conditions or operator differences.

- It is important to standardize the minimum experimental details including sample selection, collection, storage and preparation along with reporting NMR parameters as proposed by the Metabolomics Standards Initiative. This enables the interrogation and comparison of NMR data as well as facilitating experimental replication.

Pre-processing:

-The final file format, which is dependent on the programs that will be used for spectral analysis, should be in accordance with the programme that will be used for further bioinformatics analysis. For example, only the files that have been generated from Chenomx must be imported into Agilent GeneSpring.

SNP analysis

- Samples of SNP-typing plate 1, 2 and 3, including the samples with unknown SNP-gender are good enough for SNP-typing purposes. Correct gender-calling of all other samples confirms correct sample-logistics. If sample information would have been available /recorded about the possibility of maternal contamination of the child-cord blood (e.g. by mother placental blood), then maternal contamination of child cord-blood as a potential cause of gender-typing could have been better evaluated. Therefore it is recommended to register the possible maternal contamination during cord blood collection.
- Sample-id logistics seems a potential risk in multi-disciplinary projects and needs strong attention. For SNP-analysis, gender-calling is an essential control for sample-ID confirmation, which makes it essential to add other gender control samples if the sample dataset is over-represented with one gender. Repeated typing of 1-2 control samples (pref. 1 female, 1 male) in every plate is essential to resolve any unexpected sample logistics. In addition, a central relational database (which is under development e.g. in WP7) using unique subject identifiers can be helpful to overcome the problem of tracing and tracking of sample identification.

RNA analysis

- Attempts were also made by ISS to extract intact mRNA from the same samples as for the SNP analysis. However this proved to be unsuccessful, as RNA was largely or fully degraded, leading to the following recommendations:
    - Directly storing the sample in several ready-to-use smaller aliquots seems to be a more appropriate solution.
    - To have good quality nucleic acids from blood samples cells, must be directly thawed in the appropriate lysis buffer

Adductomics

Initial results that both approaches are reproducible and provide preliminary data that indicates humans are exposed to multiple classes of genotoxins . However,

- Optimisation of DNA extraction and digestion procedures are necessary to avoid artefactual loss or gain of DNA adducts (as highlighted in a previous HEALS report)
- Use of DNA, deoxyribonucleoside and base standards, is necessary to ensure accurate detection and quantitation of DNA adducts.

- Complimentary LC-HRMS and LC-MS/MS approaches increase ability to detect targeted with confidence adducts.
- Both adductomic approaches are currently relatively insensitive requiring relatively large amounts of DNA and hence further work is required to increase sensitivity or otherwise their use will be limited to tissues where large amounts of DNA are readily available.

Data fusion for omics analysis

NMR and LC-HRMS data sets produced in this work package on existing cohort samples have been treated uniquely in separate work flows for pre- processing and subsequent chemometric analytics. Fusing the data sets may provide more significant and illuminating outcomes, yielding complementary information greater than the sum of two separate approaches. For example, Forshed *et al.* In (Forshed, et al., 2007; Forshed, et al., 2007) have pioneered this approach.

Standardizing post data acquisition data analysis

Lastly, another suggestion could be to ensure that prior to multiorgan comparison of metabolome data (e.g. cord blood and maternal urine) the post data acquisition data analysis is standardized as much as possible (e.g. using the same deconvolution software, using the same reference databases to infer putative metabolite identifications). Definitely, combining the metabolome data into one paper (e.g. urine and plasma) will raise similar questions by reviewers. It does not make sense to compare e.g. metabolites from urine vs metabolites whenever these are inferred from different reference databases and draw biological conclusions from this.

4 References

Alonso, A., Marsal, S. and Julia, A. (2015) Analytical methods in untargeted metabolomics: state of the art in 2015, *Frontiers in bioengineering and biotechnology*, **3**, 23.

Amiot, A.*, et al.* (2015) (1)H NMR Spectroscopy of Fecal Extracts Enables Detection of Advanced Colorectal Neoplasia, *Journal of proteome research*, **14**, 3871-3881.

Anderson, N.L. and Anderson, N.G. (2002) The human plasma proteome: history, character, and diagnostic prospects, *Mol Cell Proteomics*, **1**, 845-867.

Badawi, AF, Mostafa MH, Aboul-Azm T, Haboubi NY, O'Connor PJ, Cooper DP. (1992) Promutagenic methylation damage in bladder DNA from patients with bladder cancer associated with schistosomiasis and from normal individuals. *Carcinogenesis* 13: 877 - 881.

Balbo S, Turesky RJ, Villalta PW. (2014) DNA adductomics. *Chem Res Toxicol* 17:356-366.

Beaudry, P.*, et al.* (2016) A Pilot Study on the Utility of Serum Metabolomics in Neuroblastoma Patients and Xenograft Models, *Pediatric blood & cancer*, **63**, 214-220.

R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, *BMC Genomics* **7**, 142 (2006)10.1186/1471-2164-7-142).

S. J. Bruce, P. Jonsson, H. Antti, O. Cloarec, J. Trygg, S. L. Marklund, T. Moritz, Evaluation of a protocol for metabolic profiling studies on human blood plasma by combined ultra-performance liquid chromatography/mass spectrometry: From extraction to data analysis. *Analytical Biochemistry* **372**, 237-249 (2008)10.1016/j.ab.2007.09.037).

Cañaveras, J.C.G. (2015) Metabolomics as a tool for the study of drug-induced hepatotoxicity. *BIOQUÍMICA y BIOLOGÍA MOLECULAR*. Facultat de CIÉNCIES BIOLÓGIQUES, pp. 284.

Carlos Cobas, J.*, et al.* (2006) A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data, *Journal of Magnetic Resonance*, **183**, 145-151.

R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, J. C. Wilson, Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems* **85**, 144-154 (2007); published online Epub2007/01/15/ (http://dx.doi.org/10.1016/j.chemolab.2006.08.014).

Emwas, A.H. (2015) The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research, *Methods in molecular biology (Clifton, N.J.)*, **1277**, 161-193.

J. Forshed, H. Idborg, S. P. Jacobsson, Evaluation of different techniques for data fusion of LC/MS and 1H-NMR. *Chemometrics and Intelligent Laboratory Systems* **85**, 102-109 (2007); published online Epub2007/01/15/ (http://dx.doi.org/10.1016/j.chemolab.2006.05.002).

R. Goodacre, D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3**, 231-241 (2007)10.1007/s11306-007-0081-3).

Jewison, T*., et al.* (2014) SMPDB 2.0: big improvements to the Small Molecule Pathway Database, *Nucleic acids research*, **42**, D478-484.

Katajamaa, M., Miettinen, J. and Oresic, M. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data, *Bioinformatics (Oxford, England)*, **22**, 634-636.

J. S. McKenzie, J. A. Donarski, J. C. Wilson, A. J. Charlton, Analysis of complex mixtures using high resolution nuclear magnetic resonance spectroscopy and chemometrics. Progress in Nuclear Magnetic Resonance Spectroscopy 59, 336-359 (2011);

Omenn, G.S., et al., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics, 2005. **5**(13): p. 3226-45.

Rai, A.J., et al., HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. Proteomics, 2005. **5**(13): p. 3262-77.

M. Rusilowicz, M. Dickinson, A. Charlton, S. O'Keefe, J. Wilson, A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics* **12**, 56 (2016); published online EpubFebruary 18 (10.1007/s11306-016-0972-2).

R. M. Salek, C. Steinbeck, M. R. Viant, R. Goodacre, W. B. Dunn, The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience* **2**, 13-13 (2013); published online Epub

Senthong P, Millington CL, Wilkinson OJ, Marriott AS, Watson AJ, Reamtong O, Eyers CE, Williams DM, Margison GP, Povey AC. (2013) The nitrosated bile acid DNA lesion O^6-carboxymethylguanine is a substrate for the human DNA repair protein O^6-methylguanine-DNA methyltransferase. *Nucleic Acids Res* 41: 3047-55.

Smolinska, A*., et al.* (2012) NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review, *Analytica chimica acta*, **750**, 82-97.

Theodoridis, G.A*., et al.* (2012) Liquid chromatography-mass spectrometry based global metabolite profiling: a review, *Analytica chimica acta*, **711**, 7-16.

Want, E.J*., et al.* (2010) Global metabolic profiling procedures for urine using UPLC-MS, *Nature protocols*, **5**, 1005-1018.

Weljie, A.M*., et al.* (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data, *Analytical chemistry*, **78**, 4430-4442.

Y. Xi, D. M. Rocke, Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. *BMC Bioinformatics* **9**, 324 (2008); published online EpubJuly 29 (10.1186/1471-2105-9-324).

J. Xia, T. C. Bjorndahl, P. Tang, D. S. Wishart, MetaboMiner – semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* **9**, 507 (2008); published online EpubNovember 28 (10.1186/1471-2105-9-507).

Zhou, B*., et al.* (2012) LC-MS-based metabolomics, *Molecular BioSystems*, **8**, 470-481.

Hebels, D.G.A.J., van Herwijnen, M.H.M., Brauers, K.J.J., de Kok, T.M.C.M., Chalkiadaki, G., Kyrtopoulos, S.A., Kleinjans, J.C.S. Elimination of heparin interference during microarray processing of fresh and biobank-archived blood samples (2014) Environmental and Molecular Mutagenesis, 55 (6), pp. 482-491.

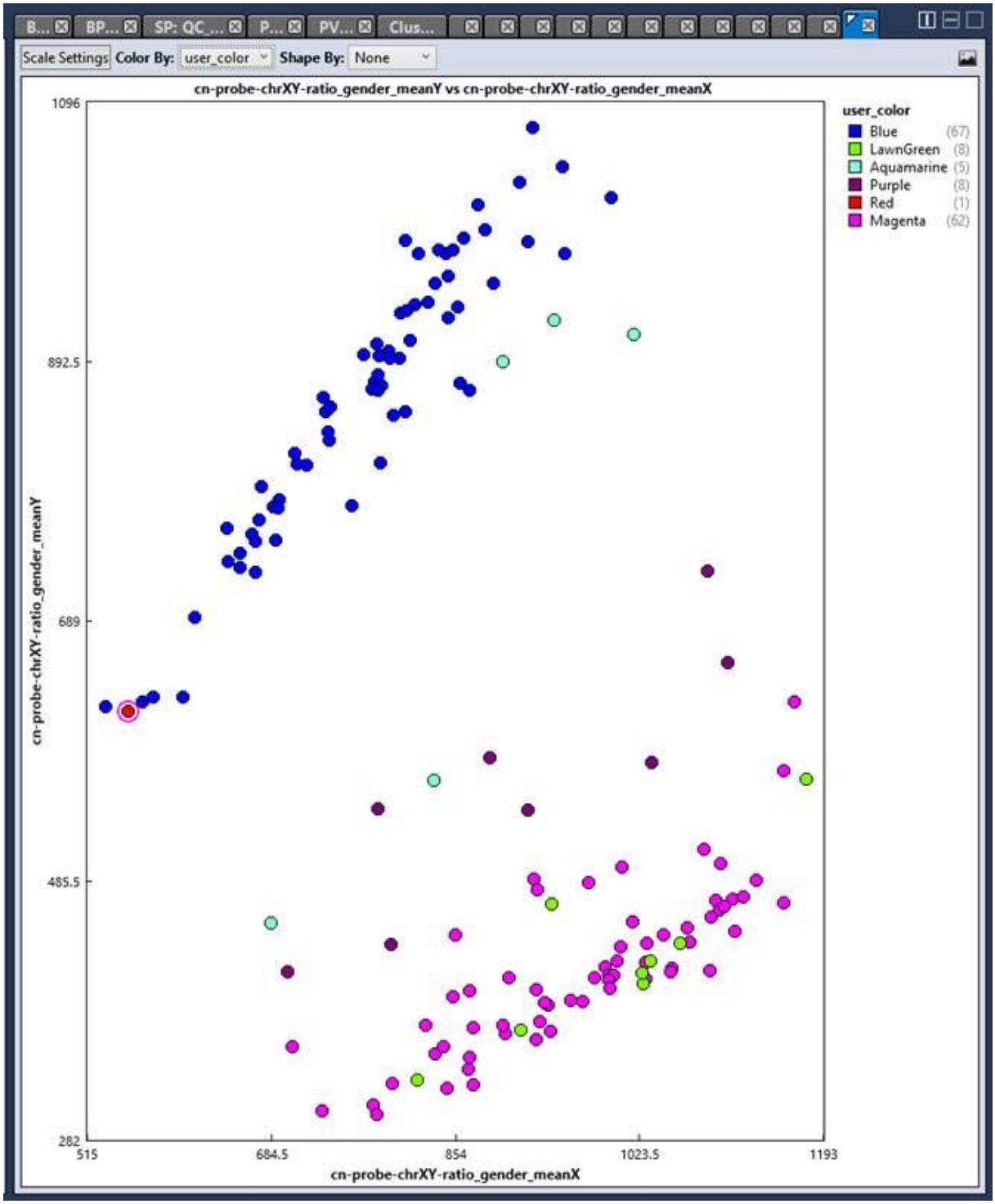
Annex 1

Fig.1: Samples with Conflicting and Unknown SNP-gender

affymetrix-plate-barcode	affymetrix-plate-peg-well position	SXS Sample Code	Customer Sample Name	DQC value	Call Rate (%)	AffyPlate	Phime / Repro-PL	Orig.Samp.Plate	Remark, GenomeScan	Repro + Phime Genders	SNP gender
p550484424894604301627 2	A04	7043-01-003-196	100097	0.9867	99.8	plate 3	Phime	Plate 3_Ingrid		F	male
p550484424894604301627 2	D08	7043-01-003-236	100164	0.9818	98.5	plate 3	Phime	Plate 3_Ingrid		F	male
p550484424894604301627 2	F12	7043-01-003-264	100209	0.9872	99.9	plate 3	Phime	Plate 3_Ingrid		F	male
p550484424894604301627 2	G11	7043-01-003-275	DD-K	0.9833	99.9	plate 3	Phime	Plate 3_Ingrid		M	female
p550484424894604301627 2	H11	7043-01-004-290	100037	0.9857	99.7	plate 3	Phime	Plate 4_Ingrid	Manually transferred	F	male
p550484424894604301627 2	C01	7043-01-004-326	R12145	0.9695	99.7	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	D09	7043-01-004-328	R12150	0.9862	99.6	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	E03	7043-01-004-330	R12153	0.9838	99.9	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	F07	7043-01-004-332	R12156	0.9754	99.6	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	H03	7043-01-004-333	R12159	0.9690	99.5	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	F	male
p550484424894604301627 2	H04	7043-01-004-334	R12182	0.9789	99.6	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	H06	7043-01-004-335	R14003	0.9789	99.8	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	H09	7043-01-004-336	R14005	0.9730	99.7	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894604301627 2	H10	7043-01-004-337	R14011	0.9671	99.5	plate 3	Repro-PL	Plate 4_Ettore	Manually transferred	M	female
p550484424894504301628 7	A02	7043-01-001-009	N21066	0.9730	97.8	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	B02	7043-01-001-010	N21089	0.9848	98.3	plate 1	Repro-PL	Plate 1_Ettore		M	unknown
p550484424894504301628 7	B03	7043-01-001-018	N21119	0.9867	98.7	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	C02	7043-01-001-011	N21101	0.9848	98.1	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	D02	7043-01-001-012	N21102	0.9798	98.6	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	E01	7043-01-001-005	N21047	0.9725	98.9	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	H02	7043-01-001-016	N21116	0.9833	99.6	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	H04	7043-01-001-032	N21157	0.9813	99.8	plate 1	Repro-PL	Plate 1_Ettore		F	unknown
p550484424894504301628 7	H06	7043-01-001-048	N21217	0.9853	99.8	plate 1	Repro-PL	Plate 1_Ettore		M	unknown
p550484424894504301629 0	F02	7043-01-002-110	R12052	0.9803	98.4	plate 2	Repro-PL	Plate 2_Ettore		M	unknown
p550484424894504301629 0	G06	7043-01-004-316	R12109	0.9754	97.7	plate 2	Repro-PL	Plate 4_Ettore	Manually transferred	M	unknown
p550484424894504301629 0	H02	7043-01-002-112	R12057	0.9823	98.2	plate 2	Repro-PL	Plate 2_Ettore		F	unknown
p550484424894504301629 0	H03	7043-01-002-120	R12087	0.9872	98.7	plate 2	Repro-PL	Plate 2_Ettore		M	unknown

Annex 2

Fig.2: ChromXY-ratio-gender plot -REPRO-PL samples: Known/snp-Gender: Blue = male/male, Lawn Green = male/female, Aquamarine = male/unknown, Purple = Female/unknown, Red = female/male, Magenta = female/female.



Annex 3

Table 1: Differences on Established Standard Operating Protocols of omics on human cohort data between D5.2 deliverables and D5.3.

	D5.2	D5.3
Instrumentation	• Metabolomics_Untargeted HR-MS, (AUTH). • Metabolomics_NMR Varian 500 MHz, (AUTH).	• Metabolomics Untargeted UPLC-MS (AUTH), ThermoFisher Scientific (Bremen, Germany) model LTQ Orbitrap Discovery MS • Metabolomics_NMR Varian 600 MHz, (AUTH).
Sample Handling	• The urine sample will be thawed immediately prior to analysis. • Centrifuge at 21000 x g for 10 minutes at 4°C. • Supernatant (200 µL) is mixed with 400 µL of chilled (4°C) methanol: water (v/v 1:1). • The samples will be analysed by Liquid chromatography – High Resolution Mass Spectrometry (LC-HRMS).	• The urine samples were be thawed immediately prior to analysis. • Centrifuged at 10000 x g for 10 minutes. • Supernatant (500 µL) is mixed with 1000 µL of chilled (4°C) methanol: water (v/v 1:1). • The samples were analysed by Liquid chromatography – Ultra Performance Liquid Chromatography (LC-UPLC).
LC conditions	• The LC column to be used is an ACE 3Q 150 x 3 mm, 3 µm (Advanced Chromatography Technologies, Aberdeen, UK) or equivalent. • Mobile phases are 0.1% formic acid in water (mobile phase A, MPA) and 0.1% formic acid in acetonitrile (mobile phase B, MPB). Gradient applied is 100% MPA for 5 minutes before increasing to 100% MPB over 15 minutes. This is held for 10 minutes before reverting back to 100% MPA and held for 2 minutes. Injection volume is 10 µl, flow rate is 0.4 mL/min and	• Acquality UPLC HSS T3 column (100 x 2.1 mm, 1.8 µm, Waters, Milford, MA, USA) maintained at a constant temperature of 40 °C. • Mobile Phase A: 100% high-grade water with 0.1% formic acid and mobile phase B: 100% acetonitrile with 0.1% formic acid or 100% methanol with 0.1% formic acid: the flow rate was 500 µl/min for the urine samples analysis in both positive and negative mode. • The following gradient was used for both positive and

	D5.2	D5.3
	column temperature set to 25°C.	negative mode: 1% B at 0 min, 1% B at 1 min, 15% B at 3 min, 50% B at 6 min, 95% B at 9 min, 95% B at 10 min, 1% B at 10.1 min and 1% B at 14 min.
MS conditions	• The MS to be used is a Thermo Exactive (Thermo Fisher Scientific, MA, USA) set at 50,000 resolution FWHM @ 200 m/z with an acquisition speed of 2 Hz, or an equivalent high resolution MS (e.g. ToF MS). Analysis will be undertaken in both positive and negative ionisation mode (separate experiments).	• The MS to be used is a ThermoFisher Scientific (Bremen, Germany) model LTQ Orbitrap Discovery MS set at 30,000 resolution FWHM @ 200 m/z with an acquisition speed of 2 Hz. Analysis were undertaken in both positive and negative ionisation mode (separate experiments).
Data analysis	• All raw data will be aligned against a QC data file before mass features are selected using either Progensis QI software (Nonlinear Dynamics, Newcastle, UK) or the open-source software XCMS (Scripps, CA, USA)..	• All raw data aligned against a QC data file before mass features are selected the open-source software MZMine v.2.21 open-software (Katajamaa, et al., 2006). • NMR data analysis proceeded using MestReNova (Mnova 11.0.3) (http://mestrelab.com/), while for the identification of metabolites it was deemed necessary to use in addition ChenoMx (http://www.chenomx.com/).

Table 2. Typical Sequence for chosen samples on REPRO analysis for both negative and positive mode

File Name	Sample ID
Test_Caffeine_1	Caffeine 20ppm
Test_Reserpine_1	Reserpine 1ppm

File Name	Sample ID
SolventBlank_1_1	Water LC Grade
QC_1	QC Samples Pool
QC_2	QC Samples Pool
QC_3	QC Samples Pool
QC_4	QC Samples Pool
QC_5	QC Samples Pool
QC_6	QC Samples Pool
QC_7	QC Samples Pool
QC_8	QC Samples Pool
QC_9	QC Samples Pool
QC_10	QC Samples Pool
R11155	R11155
R11055	R11055
R11143	R11143
R11092	R11092
QC_10-1	QC Samples Pool
R11102	R11102
R11098	R11098
R11164	R111164
Test_Reserpine_1_2	Reserpine 1ppm
SolventBlank_1_1_2	Water LC Grade
QC_1_1	QC Samples Pool
QC_1_2	QC Samples Pool
QC_1_3	QC Samples Pool
QC_1_4	QC Samples Pool
QC_1_5	QC Samples Pool
QC_11	QC Samples Pool
Test_Reserpine_100ppm	Reserpine 100ppm
R11126	R11126
N21020	N21020
SolventBlank_1_1_3	Water LC Grade
N21021	N21021

File Name	Sample ID
N21030	N21030
N21047	N21047
N21049	N21049
QC_12	QC Samples Pool
N21183	N21183
N21102	N21102
N211212	N211212
N21107	N21107
N211213	N211213
N21127	N21127
SolventBlank_1_2	Water LCMS Grade
QC_13	QC Samples Pool

Table 3.Typical Sequence for chosen samples on PHIME analysis *(SM1 code refers to Slovenia Mother sample 1from the original table has been sent) All codes were change for the convenience of the instrumentation software.

Sample Type	File Name
Blank	Caffeine_Pos
Blank	Reserpine_Pos
Blank	Water_Pos_1
QC	QC_Pos_1
QC	QC_Pos_2
QC	QC_Pos_3
QC	QC_Pos_4
QC	QC_Pos_5
QC	QC_Pos_6
QC	QC_Pos_7
QC	QC_Pos_8
QC	QC_Pos_9
QC	QC_Pos_10
QC	QC_Pos_11
Unknown	SM_1_Pos

Sample Type	File Name
Unknown	SM_2_Pos
Unknown	SM_3_Pos
Unknown	SM_4_Pos
Unknown	SM_5_Pos
Unknown	SM_6_Pos
Unknown	SM_7_Pos
Unknown	SM_8_Pos
Unknown	SM_9_Pos
Unknown	SM_10_Pos
QC	QC_Pos_12
Unknown	SM_11_Pos
Unknown	SM_12_Pos
Unknown	SM_13_Pos
Unknown	SM_14_Pos
Unknown	SM_16_Pos
Unknown	SM_17_Pos
Unknown	SM_18_Pos
Unknown	SM_19_Pos
Unknown	SM_20_Pos
QC	QC_Pos_13
Blank	Water_Pos_2

Annex 4

Table 4. MZmine parameters for data processing for REPRO and PHIME cohort study.

	Urine	
	ESI (+)	ESI (-)
Baseline correction		
Type	base peak chromatogram	base peak chromatogram
Smoothing	500	500
Asymmetry	0.001	0.001
m/z bin width	1	1
Mass detection		
Algorithm	centroid	centroid
Noise level (cps)	1E6	7E5
Chromatogram builder		
Minimum time spam (min)	0.05	0.05
Minimum height (cps)	1E6	7E5
m/z tolerance (ppm)	5	5
Chromatogram deconvolution		
Algorithm	Wavelets (XCMS)	Wavelets (XCMS)
S/N threshold	5	5
Wavelet scales (min)	0.1-5	0.1-6
Peak duration range (min)	0.1-1	0.1-1

Deisotoping		
Algorithm	Isotopic peaks grouper	Isotopic peaks grouper
m/z tolerance (ppm)	5	5
t_R tolerance (min)	0.05	0.05
Identification of adducts	√	√
Identification of peak complexes	√	√
Alignment		
Algorithm	RANSAC	RANSAC
m/z tolerance (ppm)	5	5
t_R tolerance (min)	0.03	0.03
t_R tolerance (min) after correction	0.05	0.05
RANSAC iterations	15000	15000
Minimum number of points (%)	20	20
Threshold	4	4
Model	non linear model	non linear model
Gap filling		
Algorithm	Peak finder	Peak finder
Intensity tolerance (%)	80	80
m/z tolerance (ppm)	5	5
t_R tolerance (min)	0.05	0.05

Annex 5

Table 5. Final outcome from REPRO_PL analysis.

Biological Fluid	Analytical Technique	Identified Peaks	Annotated Peaks	Unique Elements (Identified Metabolites)
Urine	LC-MS (negative mode)	3239	1715	527
	LC-MS (positive mode)	1009	623	103
	NMR	40	40	40
Plasma	LC-MS (negative mode)	267	264	100
	LC-MS (positive mode)	430	380	128
	NMR	49	49	49

Table 6. Final outcome from PHIME analysis.

Samples	Biological Fluid	Analytical Technique	Identified Peaks	Annotated Peaks (Metabolites)	Unique Elements (Identified Metabolites)
Mothers	Urine	LC-MS (negative mode)	3239	1715	99
		LC-MS (positive mode)	1009	623	209
Children		LC-MS (negative mode)	267	264	275
		LC-MS (positive mode)	430	380	181
Mothers	Plasma	LC-MS (negative mode)	2501	612	611
		LC-MS (positive mode)	4497	1672	1671
Children		LC-MS (negative mode)	2918	807	807

		LC-MS (positive mode)	7662	2522	2522

"This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement N°603946